Technical Report: A Simulation Study of Hadoop for Hybrid Networks (HHN)

Xiaoyu Wang
University of Virginia
Charlottesville, VA 22904–4743
xw5ce@virginia.edu

Malathi Veeraraghavan
University of Virginia
Charlottesville, VA 22904–4743
mv5g@virginia.edu

I. Workloads

Workloads: We started with the Facebook 2010 (FB-2010) workload [1], which provides the following information for each job: (i) arrival time instant, (ii) input dataset size, (iii) shuffle data size, (iv) reduce output size, (v) number of map tasks, and (vi) number of reduce tasks. Using an assumption that each map task processes one input block of size 128 MB, and that the number of reduce tasks is equal to the number of map tasks divided by 8, we derived the number of map tasks and number of reduce tasks of each job from the size of its input dataset. As no information was provided about task execution times, we chose 15 s and 25 s for each map task and reduce task, respectively.

TABLE I: Trace sets composition

No. of maps	% regular jobs	
	RJS1	RJS2
1-9	40%	20%
10-99	40%	50%
100-499	18%	28%
500-10000	2%	2%
Shuffle-data size		
0	10%	
0-0.8 GB	70%	
0-2 GB	20%	
	TS1	TS2
Regular jobs	RJS1	RJS2
SHJs	first 40 SHJs of FB-2010 workload	first 60 SHJs of FB-2010 workload with input-data size smaller than 800 GB

In the FB-2010 workload, more than 50% of the jobs are very small jobs, with only one or two map tasks. With the given job arrival times, the original workload results in low CPU utilization, i.e., around 10%, in our simulation. To achieve higher CPU utilization levels, we generated *two larger Regular-Job Sets (RJS)*, and named them RJS1 and RJS2. Table I shows the job-size and shuffle-data-size distributions of RJS1 and RJS2. To determine the characteristics of a regular job, we first generated two samples between 0 and 1 assuming the uniform distribution. The first sample was used to determine the job-size range. For example, if the sample was 0.45, the job-size range is 10-90 map tasks in RJS1 (see Table I). The second sample was used to select the number of map tasks for the job from within the range 10-90. Similarly, a random uniformly-distributed sample was drawn to select the shuffle-data size using the percentages shown in Table I. We defined jobs with a shuffle-data size larger than 2 GB as shuffle-heavy (SH) jobs.

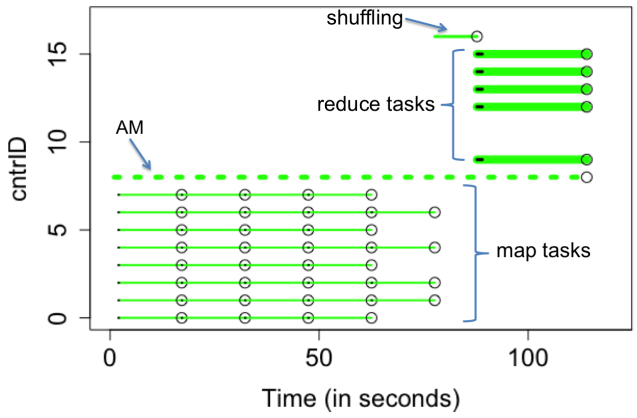
The next step was to create two *Trace Sets, TS1 and TS2*, which mix SH jobs with regular jobs from RJS1 and RJS2, respectively. As shown in Table I, specific SH jobs were drawn from the FB-2010 workload for the two trace sets TS1 and TS2. The reason of removing larger SHJs in TS2 is that early simulation runs showed that one very large SHJ could skew the results, as shown in Section IV-C of our paper [2].

For each simulation run, a different trace of jobs was created using the following approach. First, an exponentially distributed random sample with parameter λ was drawn for the inter-arrival time to the next job. Next, a parameter called *SH-job percentage* p_s was used to set the percentage of SH jobs in a trace. For each job, a Bernoulli distributed sample with parameter p_s was drawn to decide whether the job should be a SH job or a regular job. If it was a SH job, then the parameters of the next SH job in the FB-2010 workload were simply selected. Thus, in all runs, job traces using the TS1 settings had the exact same 40 SH jobs, and these jobs appeared in the exact same order. Similarly, all SH jobs in traces based on the TS2 settings had the exact same 60 SH jobs, and these jobs also appeared in the exact same order. The exact positions of these 40 or 60 SHJs within the trace varied from run to run because of the p_s based random sampling for choice of job type.

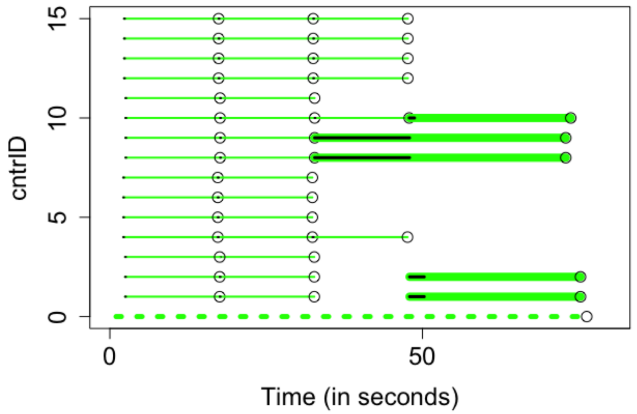
II. Simulation Results

A. Effect of clumping on a single shuffle-heavy job

Using the modified data-block placement policy, the input dataset of a shuffle-heavy job is concentrated to a few number of racks, which limits the amount of computing resources accessible for the job. To study the effect of input-data clumping, we start with a single shuffle-heavy job in a small system. We simulate a cluster of two racks, in which there are a total of 16 containers indexed from 0 to 15. The shuffle-heavy job consists of 36 map tasks and 5 reduce tasks. In the hybrid architecture the input dataset is stored only in the first rack, while the dataset is stored in both racks in the EPS-only architecture. The optical link rate is 5 Gbps in OSM.



(a) Modified Hadoop in OSM



(b) Original Hadoop in the electrical architecture

Fig. 1: Start and finish time of map tasks, reduce tasks and shuffling of a single shuffle-heavy job

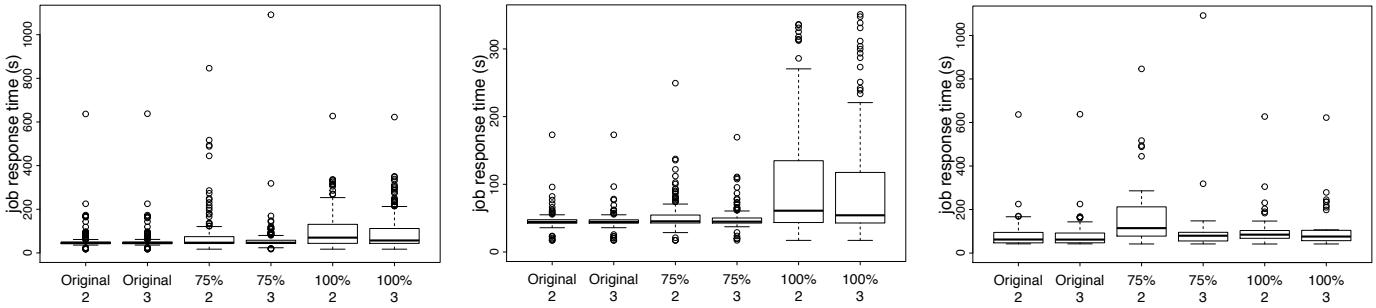
The inter-rack electrical link rate is 500 Gbps in the hybrid architecture and 5.5 Gbps in the EPS-only architecture.

Fig. 3 illustrates how containers are allocated to the shuffle-heavy job when it runs in the two architectures. The dashed line represents the AM container. The thin and bold lines in green correspond to map-task containers and reduce-task containers, respectively. The black segments shows the time period when map output is being shuffled. The job is completed faster in the electrical architecture than in OSM (75 s vs. 113 s). This is because the job can only use the 8 containers in the first rack to execute map tasks due to its concentrated input dataset in OSM, while it can use all the 15 container (except for container 0 used by the AM) to execute map tasks. On the other hand, thanks to the decoupled shuffle phase from reduce tasks, reduce containers do not need to sit idle when waiting for the shuffle phase to finish with the modified Hadoop.

B. Effect of number of replicas

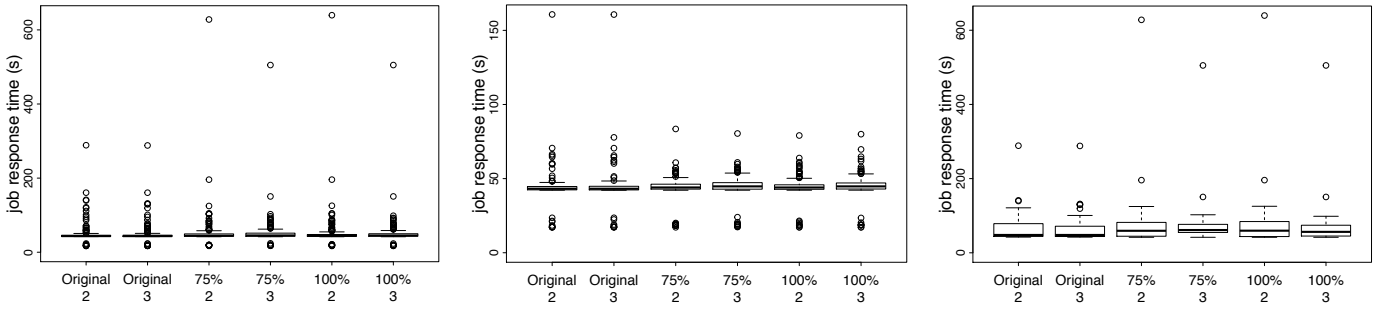
Fig. 1 and Fig. 2 show boxplots to study the effect of the number of dataset replicas on job response time in a 4-rack system and a 12-rack system, respectively. The traces used in these runs belong to TS1.

We make the following observations: (i) Job response time is not notably affected by the number of replicas for the original Hadoop on EPS-only network. (ii) In the smaller system with 4 racks, both regular jobs and SH jobs could benefit from a larger number of replicas (i.e., 3 replicas) in HHN. (iii) In the larger system with 12 racks, response times of SH jobs are reduced when having 3 replicas compared to 2 replicas, while response times of regular jobs almost remain the same for different replica values.



(a) All jobs (b) Regular jobs (c) Shuffle-heavy jobs

Fig. 2: Job response time comparison in a 4-rack system; 2 and 3: number of replicas; Original: original Hadoop on EPS-only network; 75% and 100%: HHN-75% and HHN-100%; TS1 input; p_s=20%; λ=0.3



(a) All jobs (b) Regular jobs (c) Shuffle-heavy jobs

Fig. 3: Job response time comparison in a 12-rack system; 2 and 3: number of replicas; Original: original Hadoop on EPS-only network; 75% and 100%: HHN-75% and HHN-100%; TS1 input; p_s=20%; λ=1.5

References

[1] "SWIM workload repository," https://github.com/SWIMProjectUCB/SWIM/wiki/Workloads-repository.

[2] X. Wang and M. Veeraraghavan, "An evaluation study of a proposed Hadoop for hybrid networks (HHN)," Submitted to ICC'2017.